Multimodal Modeling and Analysis of Intelligent Systems

A

synopsis

submitted in the partial fulfillment for the requirement of

Doctor of Philosophy
in
Computer Science

submitted by

Dhruv Bhandari

under the supervision of

Prof. Sandeep Paul
Supervisor
Professor
Dept. of Physics & Computer Science,
Dayalbagh Educational Institute

Prof. Huzur Saran
Co-supervisor
Professor
Computer Science & Engineering,
IIT Delhi

DEPARTMENT OF PHYSICS & COMPUTER SCIENCE
DAYALBAGH EDUCATIONAL INSTITUTE
(DEEMED UNIVERSITY)
DAYALBAGH, AGRA UP 282 005
2020

Introduction

The human interpretation of the surrounding environment is intrinsically based on our perception capabilities. Multimodal yet unified nature of human perception plays a central role in the development of human intelligence(Quek et al., 2002; Smith and Gasser, 2005). We regularly rely on our senses to continuously learn new information by interacting with other humans or in some cases with our surrounding environment. These interactions could be active for instance when we are discussing with our partner or passive when we are listening to a teacher or reading a book written by another human. In general, we often use a combination of them to experience an external stimulus such that the knowledge gained is treated as a unitary representation in the development of our intelligence.

With the advancements in machine learning in the last decade, a sincere effort is being made to gradually inch closer to human level machine intelligence in various fields like computer vision, pattern recognition, language models etc. But the pioneers in this field still argue that human level machine intelligence is possible but only in the far future (Architects of Intelligence).

In order for machine intelligence to make progress in understanding the world around us, it needs to be able to interpret and reason multimodal messages. The focus has to shift from using processing unimodal signals in current machine learning setups to building models that can process and relate information from multiple modalities (Baltrušaitis et al., 2019).

The advent of deep learning has proven promising for multimodal machine learning(Chao et al., 2015; He et al., 2015; Ngiam et al., 2011; Uzan and Wolf, 2015; Wang et al., 2014; Wu et al., 2016; Zhang et al., 2016).It has also proven to be very effective in making highly accurate predictions on huge and complex data from multiple sources. These predictions are generally based on the task driven and computationally intensive learning of deep neural network models on the input data. It can be argued that, while these approaches are powerful and accurate, they do not offer suitable explanation on how learning is happening. The knowledge embedded in the system is distributed over the connections between the neurons in the form of numbers. In the big data era, tons of data, usually multimodal in

nature (audio, video, speech, text etc) is being generated in all kinds of industrial and scientific environment around the world. However, with the data quantity increasing, another critical issue of data ambiguity which involves noise and uncertainty emerges as a challenge to machine learning research community. Therefore, it becomes evident that we use robust fusion techniques and algorithms that are able to give logical explanations for the learning mechanism that these networks follow.

Motivation

Researchers have been discussing the relationship between top-down information processing and bottom up information processing in humans (Architects of Intelligence). When one looks at an image, the light falls on their retina and they interpret the object in the image in a bottom-up manner whereas when they use their experience of how objects in the image behave to add perception to that image, they process it in a top-down manner. Machine learning systems lack this perceptual capability and only capture one side of the equation – the bottom-up side. They need to be able to capture both sides of that equation for developing common sense reasoning like humans. Thus, the multimodal nature of human perception is an inspiration for making machine learning models process information in the top-down manner along with the existing bottom-up manner. Modality in its true sense is the representation format for exchange of Information/Communication. When a combination of them is used to represent or exchange information, it is referred to as multimodal representation.

As an extension to this school of thought, multimodal machine learning has enabled a wide range of research applications. A solution to the above-mentioned challenges will help in application areas like:

Affective Computing: Affective Computing can be considered to be an umbrella for sub application domains like:

- Emotion Recognition,
- Sentiment Analysis,
- Personality or Trait Recognition in terms of human behavior.

Emotion Recognition is the process of identifying human emotions in terms of categorical labels like happiness, sadness, anger etc or dimensional labels like arousal, valence etc. The field usually takes into account multiple modalities like image (Facial expressions), Speech/Text (voice), Physiological signals like EEG, ECG etc. to identify emotions at any given time. Although recognizing emotions is easier for humans, it can become challenging when it comes to intelligent systems. This is predominantly due to various modalities through which human emotions can be expressed, the abstraction levels at which information from them can be fused, the sophisticated methods that are employed for fusing them and lack of temporal boundaries separating various emotions. A large percentage of the emotion recognition research focuses on grouping emotions in discrete categories, whereas, there are others that take help of a dimensional approach for describing emotions. The later approach has certain advantages over the former. First being that it can predict time continuous labels which results in more realistic emotion recognition. Second is with regards to representing a broader spectrum of emotions when compared with Discrete or Categorical case. The advancements in human emotion recognition techniques will result in a gradual shift in human computer interactions. It will become more users centric rather than being computer centric and in turn improve the communication of information through affective display of emotions. This makes human emotion recognition one of the most researched topics in affective computing community.

Sentiment Analysis on the other hand, can be defined as the process of computationally identifying and categorizing opinions expressed by a human about a particular topic or a product. It usually deals with classifying the opinions into discrete classes – positive, negative or neutral. Sentiment Analysis is fundamentally different from classification of media into various categories since it requires a higher level abstraction to understand the message in the media content. The primary advantage of using multimodal approach over the unimodal approach for sentiment analysis (text based, as is predominantly used in the recent past) is the abundance of behavioral cues. With social media platforms, digital media, multimedia content becoming more prominent, people are extensively using multimodal approaches to express their opinions using visual cues (You Tube, Vimeo etc.) and audio signals (Podcasts). Therefore, it becomes increasingly crucial to mine opinions from these diverse multimodal

information sources. Understanding these opinions is no longer confined to a single modality but calls for a cross-modal approach of the same.

Similarly, Personality or Trait Recognition deals with classifying human behavior into Big 5 personality traits – openness, conscientiousness, extraversion, agreeableness and neuroticism.

Media Description/Information Retrieval: In the big data era, huge amount of multimedia data is being generated from various sources like social networking websites, news articles, documents that are available online and advertisements etc. Most of them are images or videos and are without any descriptions. With our knowledge and prior experience, we can easily understand them without detailed captioning. However, for machines it becomes difficult to interpret them and generate well formed descriptions. This indirectly leads to the failure of machines in retrieving information about the object in the image or video.

The research in Media Description and Multimedia Information retrieval could be further sub-divided into application areas like:

- Image Captioning,
- Video Captioning,
- Visual Question Answering,
- Cross Modal Information Retrieval

It lies at an intersection of Computer Vision and Natural Language Processing. The task in Image Captioning and Video Captioning is to provide a free form text description of a piece of media (Image or Video Clip). Similarly, for visual question answering a cue image or video is given and particular questions having answers in the form of yes/no are expected. The free form of question answering from this visual cue is still not mature and requires highly accurate media captioning (Image or Video) to better the results.

Event Recognition: Event Recognition can be further sub-divided into three categories:

- Human Action Recognition
- Human Gesture Recognition
- Segmentation of scenes

The focus here is to detect a predefined event or scene given an audio, video or text input. Since human action is usually a sequence of events occurring one after the other, in human action recognition or gesture recognition the most important question that needs to be answered is how to deal with the temporal information? In this context, action and gesture recognition are challenging problems due to the following reasons –

First, the amount of data that has to be processed. The difficulty here is handling the tradeoff between spatial information processing and temporal information processing. Often while processing such large quantities of data, either a frame sub sampling method is taken, or the local frame level features are aggregated into mid level video features. But both of these methods have their limitations when it comes to critical applications like video surveillance, anomaly event detection in a mob/crowd etc. A more recent approach of temporal sequence modeling has been used to take advantage of spatiotemporal features.

Second challenge comes up with fusion methodologies. As described above, human action/gesture is usually a sequence of events, which use multimodal information to process spatiotemporal information and require simultaneous fusion across both special as well as temporal features.

The third and the final challenge that comes up in this task is related to model complexity. With huge amount of information being processed both in spatial as well as temporal dimensions, the models become bulky and also difficult to tune/train. Another challenge is the limited availability of labeled dataset, which forces research to focus more towards unsupervised learning approaches. The above mentioned challenges in Event Detection and its critical application in areas like security video surveillance, mob activity detection through anomalous behaviors, Human Computer Interaction make it a potential research problem for multimodal machine learning community.

With this discussion on the application domains of multimodal intelligent systems along with existing challenges in them, a more detailed analysis of each of these domains is required. The next section elaborates existing techniques, limitations, their proposed possible solutions, existing datasets and drawbacks for each of those mentioned above.

Review -Applications of Multimodal Modeling

The abundance of multimodal information present in the surrounding environment raises an obvious requirement in the development of intelligent systems - to be able to process multimodal information and simulate human perception. This is necessary to bridge the adaptive gap between humans and machines. Moreover, the challenges that have been detailed above in the each sub section may not have an exclusive one to one solution for each of them – an integrated approach may be required for solving them. Even though this problem of multimodal modeling for intelligent systems may seem challenging at the outset, multiple attempts have been made to solve a subset of challenges in the below mentioned approaches:

Affective Computing: The focus of multiple research papers on multimodal emotion recognition has been visual and aural information. Annual competitions like Audio Video Emotion Recognition Challenge (AVEC) and EmotiW have encouraged benchmarking Multimodal Emotion Recognition using deep learning approaches(Chao et al., 2015; He et al., 2015; Povolny et al., 2012; Rouast et al., 2019; Valstar et al., 2016; Wang et al., 2014).One such Deep Learning model is proposed by Zhang et.al in (Zhang et al., 2016). The authors use a two folds approach – First, two Deep Convolution Neural Networks are pre-trained on a large scale image data and then fine tuned to carry out feature extraction part on the corresponding labeled speech and face data. In the second phase, a fully connected network is used to fuse features from the above two Deep Convolution Neural Networks. Emotion recognition task is then carried out on the result of the above mentioned fusion model. This was one of the very first works on fusing multiple modalities using deep learning. Another such robust multimodal deep neural network model is proposed in (Wu et al., 2016). The authors illustrate a unified and hybrid approach to

speech enhancement which is inspired by insights from human audiovisual perception. A Convolution Neural Network is used to extract useful features from visual cues where as a fully connected Deep Neural Network is used for identifying features from audio signals. They are then integrated using a Bidirectional Long Short Term Memory (BiLSTM) Network. The authors also indicate that their hybrid approach to training learns to adjust the importance of the audio and visual streams intrinsically based on the uncertainty in audio stream. The parameters of the resulting hybrid network are then tuned using back propagation. The results in the paper also demonstrate the superiority of such hybrid approaches in the noisy signal environment. In (Chao et al., 2015) the authors use Long Short Term Memory Recurrent Neural Network (LSTM-RNN) for dimensional emotional recognition. An intuitive loss function is used here to optimize in comparison to the mean squared loss function. They prove through their results that the intuitive loss is more robust for label noises and it can ignore small errors to get a stronger correlation between predictions and labels. Their work also proposes a temporal pooling technique which is similar to the pooling in a convolution neural network and pools the features from temporal inputs rather than spatial inputs. The temporal pooling approach helps in achieving superior performance by getting the statics of successive frames and by increasing the diversity of features fed to the subsequent layers.

However, even with multimodal emotion recognition achieving benchmark performances on various datasets, it still has limitations when it comes to its applications in human-computer interactions (HCI). The humans use their raw multimodal perception and are at an adaptive advantage since this approach reduces the amount of information that has to be processed to more manageable proportions(Lewkowicz, 2000). Moreover, the temporal characteristics of these affective/emotional states have huge variations since each individual may express emotions differently. The human perception comes into play in such situations. Most of the existing intelligent systems are still inefficient at these perceptual tasks. Therefore, it becomes increasingly evident to bridge the adaptive gap mentioned above and focus on building models which take perception into account while solving emotion recognition problems.

Similarly, multimodal sentiment analysis proves to be of utmost importance with the ever growing multimedia data present on the internet, there is scarce literature present in this field (Mihalcea et al., 2013; Poria et al., 2015, 2018; Poria, Cambria, Bajpai, et al., 2017; Zadeh, 2015). Most of the work in sentiment analysis has been carried out in the field of natural language processing and limited effort has been made to incorporate other modalities (Borth et al., 2013; Chikersal et al., 2015; Kim, 2014; Kim et al., 2015; Richard Socher, Alex Perelygin, Jean Y.Wu, Jason Chuang, Christopher D. Manning, 2013; You et al., 2015). However, with social media platforms becoming more prominent, people are extensively using multimodal approaches to express their opinions using visual cues (You Tube, Vimeo etc.) and audio signals (Podcasts). Therefore, it becomes increasingly crucial to mine opinions from these diverse multimodal information sources. One such effort towards visual sentiment analysis has been made in (Borth et al., 2013). The authors propose a novel approach which tries to understand visual concepts that are strongly related to sentiments. The paper highlights two key contributions by the author: First, A large scale visual sentiment ontology which consists of Adjective Noun Pairs (ANP) is constructed. It takes inspiration from various psychological theories present in the literature. Second, the authors propose SentiBank, a novel visual concept detector which can be used to detect the presence of such Adjective Noun Pairs in an image. The paper suggests that use of such visual models along with ANP can improve detection of sentiments over text-based models while experimenting on image tweets. Inspired by such models, You et.al in (You et al., 2015) proposed a Convolution Neural Network trained to perform Image Sentiment Analysis on Flickr images. Further, they improve the performance on twitter images by inducing domain transfer with a small quantity of manually labeled twitter images. The model is evaluated on this manually labeled dataset and the results show that CNNs achieve better performance on image sentiment analysis than competing algorithms. An attempt to use audio, speech and visual cues in addition to the textual cues was made in (Wollmer et al., 2013). Another attempt to use the state of the art LSTM to grasp the context evidence from surrounding utterances for multimodal sentiment analysis was made in (Poria, Majumder, et al., 2017). Their results showed significant improvement in this field. Inspired by these improvements, the authors extended their work by fusing multiple modalities using an attention based approach in (Poria, Cambria and Hazarika, 2017).

Even though a number of recent approaches to multimodal sentiment analysis have produced interesting results, there are major issues which are still unaddressed. These include consideration of context while classification, effect of speaker-inclusive and speaker exclusive scenarios, the impact of each modality across datasets and generalization ability of multimodal sentiment classifier. A recent study (Poria et al., 2018)proposes methods to handle some of these issues like speaker independence and generalization of the models. The work from the authors in (Poria et al., 2018)is important in the context of progress of multimodal sentiment analysis since it also performs behavioral studies on their methods for solving some of the above mentioned problems. One of the behavioral studies is based on three different fusion techniques, each improving on performance in comparison to the previous one – Concatenation of features, Context aware fusion, and Context Aware Fusion with attention. In another behavioral study in this paper, the authors experimented by training on one dataset and testing on other similar kind of dataset for handling the generalization problem. However, even with such studies, there is a huge scope of improvement in this field by solving the above-mentioned issues – it still remains an open domain for research work in Affective Community.

Multiple multimodal benchmark dataset for affective computing have been proposed for the above-mentioned sub domains. For emotion recognition AFEW – Acted Facial Expressions in the Wild (Dhall et al., 2011), RECOLA – Remote Collaborative and Affective Interactions (Ringeval et al., 2013, 2015; Valstar et al., 2016) have been used. For Sentiment analysis, POM – Persuasive Opinion Multimedia (Park et al., 2014), MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (Zadeh et al., 2016) have been proposed.

To sum up, a common ground in the above-mentioned approaches of affective computing is that first unimodal features are extracted and then either feature level, decision level or a hybrid fusion approach is adopted for integrating information from all modalities. Unimodal feature extraction systems along with the fusion systems form the basic building blocks for a well performing multimodal system. This makes multimodal research in affective computing an open and challenging domain for researchers.

However, it finds uses are across various important domains like healthcare, security, social engineering etc. which is why even a trivial contribution to this field becomes hugely significant.

Media Description/Information Retrieval: Much like the machine translation models, where a recurrent network is first used to identify a good representation of the input text (Encoder RNN), and then this representation is feed to a second RNN (Decoder RNN) for generating the translation, in Media Description models, the encoder RNN is replaced by a Convolution Neural Network which maps the Input image to a visual space. The decoder RNNs directly generates the output translation from this latent visual space (Vinyals et al., 2014).The information regarding the image/video and the corresponding caption is directly passed into the RNN decoder. In contrast to this approach, a joint representation can also be learnt in the multimodal space instead of a visual space (Karpathy et al., 2014; Kiros et al., 2014a, 2014b). The focus here is to extract features from the image/video and text modality in the multimodal space and then generate language (words/sentences) from this information.

Although the performance of such methods is improving continuously from translation-based methods to many attention-based methods, there are issues with these approaches as they are supervised in nature and usually require a large dataset for training. Moreover, since the text/caption generation network's capability depends on how well features are extracted by the encoder network, the application is limited to a domain specific dataset. Multiple unsupervised (like General Adversarial Networks - GANs, Conditional General Adversarial Networks) and reinforcement based techniques have been proposed which are better suited for this task since they can create a diverse set of image captions in contrast to the conventional supervised method explained above(Dai et al., 2017; Ren et al., 2017; Rennie et al., 2017; Shetty et al., 2017).An attempt to address the diverse space problem was also made in (Niu et al., 2018). The authors employ a multi-scale CNN for image annotation to extract features and also fuse them at different levels of abstraction. The supervised methods are trained with an objective to reproduce the captions as provided by the ground truth and hence lack the diversity in the captions they generate. The recurrent models (in supervised techniques) which are generally used in text generation predict a word conditioned on an image and all previous words. Consequently, if the previous words

generated were incorrect these models become more susceptible to error. Although GANs have improved the performance of image/video captioning, Visual Question generation systems, they are nowhere close to the diversity which is reflected by human ground truths/captions. Since the information in an image/video varies across multiple regions of itself, the Generative models with attention would prove to be an interesting area of research.

Multiple benchmark datasets for media description have been proposed. Image Captioning has been benchmarked in CVPR 2015 by using the Microsoft Common Objects in Context (MS COCO) dataset (Lin et al., 2014). For Video Description/Captioning, MPII Movie Description dataset and Montreal Video Annotation dataset have been used (Rohrbach et al., 2015, 2017; Torabi et al., 2015).

Event Recognition: In this context the domains of human action recognition and gesture recognition can be further categorized into three groups: 3D Convolution based methods which focus on spatio-temporal learning structure with time (frame sequence) being the third dimension (Baccouche et al., 2011; Ji et al., 2013; Varol et al., 2018), Motion based which take into account dense optical flow like methods before feeding the input to a CNN(Varol et al., 2018; Weinzaepfel et al., 2015) or Temporal method based which use sequence models like RNN or LSTM along with CNNs (Mahasseni and Todorovic, 2016; Singh et al., 2016; Veeriah et al., 2015).

The 3D convolution based methods face problems when video sequences are too long since the kernel in the third dimension becomes large, increasing the number of tunable parameters. This also requires them to train on very large datasets. In contrast to 3D CNNs, Temporal models focus more towards keeping the information in the frames intact for a longer period of time. A video summarization approach has been discussed in detail in (Rochan et al., 2018). Such methods can also be used for segmentation of scenes and summarization of videos. An interesting application of event recognition is in education domain where bullet points of topics covered in discussions/lectures can be produced for students to identify the key highlights in them. Another challenging but interesting task is to search where in the entire catalog of video lectures, a specific topic is being discussed.

The Motion based models have recently found new research avenue in Human Action Prediction. It is relatively difficult when compared to Action Recognition tasks since the latter is an after the video analysis format. While in Action Prediction, the task is to identify from the set of current frames what the activity is likely to be. In addition to the above mentioned challenges in this application domain,(Kong and Fu, 2018) introduces multiple other challenges – Inter vs intra class variation, Cluttered Background and Camera Motion while capturing and Scarcity of annotated data which makes this a prospective research field.

Preliminary Work Done

Human Perception, due to its unified nature plays a strong role in our interaction with the surrounding environment. There is abundance of multimodal information present in the environment. Therefore, humans have been trained to process this multimodal information and treat it as a unitary representation of the world.

Inspired by this multimodal yet unified nature of human perception, an attempt is made in (Bhandari and Paul, 2018)to move towards robust intelligent machines which have the ability to process multiple modalities at once and treat the information from them as a unitary event. Deep Neural Networks (DNN) have been employed and the Affect/Emotion Recognition task has been explored exhaustively. Three DNN models are proposed in this paper – Unimodal Perception Net, Multimodal Perception Net and Hierarchical Multimodal Perception Net. Each of them improves on the performance in affect recognition task when compared with the previous one. They are designed around focal theories of human perception – Multisensory Enhancement Theory or Super-additive Effect of multisensory integration and Principle of Inverse Effectiveness. The Multisensory Enhancement Theory states that the response to multimodal stimuli is more than that of independent unimodal stimuli taken together. As an extension to this, the Principle of Inverse Effectiveness states – if the response to each component is weak, then the enhancement opportunity is much more and vice versa. Efficient hierarchical strategies

for facilitating multimodal data fusion are also illustrated in this work. These strategies reduce the number of tunable parameters hence making the models easier to train.

The performance of these models is tested on a benchmark multimodal dataset - RECOLA (Remote Collaborative and Affective Interactions). The results show that these models successfully fuse modalities which supplement each other and hence give insights to why a particular combination of modalities is better suited for identifying emotions. Their performance is significantly accurate when compared to other existing techniques; even if a smaller number of modalities are used

However, even with significantly accurate performance on complex affective recognition tasks, these models lack the ability to tap amodal information like rhythm, tempo and tones. Humans on the other hand are extremely efficient in doing so. This is illustrated in the third focal theory of human perception – the inter-sensory redundancy hypothesis. An extension to this work can include development of models and approaches which are equipped with the abilities to handle both modal as well amodal information.

Proposed Methodology And Discussion on Multimodal Modeling and Analysis

The above mentioned review of various multimodal modeling applications allows us to identify the existing limitations and motivates us to work towards the following new research avenues. The multimodal signals used in applications mentioned above show diverse characteristics. To handle this diversity, it is proposed that as an initial step, the following technical aspects should be taken care off:

- Representation: The heterogeneity of multimodal signals makes representation of multimodal data difficult. The models have to exploit the complementary nature of the multimodal signals as well as take care of the redundancy in the data.
- Translation: The mapping of data from one modality to the other is difficult since the translation relationship may be subjective since it may be a real valued signal in case of

speeches or symbols in case of sentences. It can be understood as generating one entity in a modality given the same entity in any other modality.

- Fusion: The concept of integrating information from multiple modalities with the aim of classification, prediction or regression.
- Alignment: Identifying the direct relationships between sub elements from two or more modalities.
- Co-Learning: Learning the modeling of one modality using the other when the former either has less annotated data, corrupt or noisy input etc.

These technical cores are the fundamentals of this study; it is proposed that as an initial step, the existing approaches in each of these technical cores of multimodal machine learning will be studied thoroughly. The aim here would be to investigate the limitations in them. The identified limitations would lay the foundation to explore possible improvements in the technical cores of multimodal machine learning so as to propel towards the either improved or novel multimodal learning algorithms.

The knowledge and improvements from the first two steps will be utilized to better understand unresolved problems in the application domains mentioned in this study and implement improved versions of multimodal machine learning algorithms to solve them.

In this context, a high-level overview of the proposed methodology is discussed below for a few application domains.

A major application domain of Multimodal Machine Learning is Affective Computing. As mentioned in the previous section, in the recent past research in sentimental analysis under affective computing is focused on using only text as input modality. With the increasing popularity of multimedia internet platforms like youtube, podcasts etc an efficient approach for Fusion, Representation and Translation needs to be identified for introducing multimodal capabilities in sentiment analysis techniques.

Similarly, with the drawbacks of supervised media captioning models, focus has now shifted towards unsupervised and reinforcement based approaches as discussed above. However, it is still very easy for anyone to differentiate between captions generated by a machine and those generated by a human. Therefore, to introduce diversity in the captions generated by these models, better and more robust Representation, Alignment and Translation strategies are needed. Sentiment Analysis and Emotion Recognition are tasks which require a higher level of abstraction for developing a deeper understanding of the message in the media content. The above mentioned approaches – Sentiment Analysis, Emotion Recognition from Affective Computing and Image/Video Captioning under Media Captioning can be integrated to come up with enhanced Intelligent Systems which have abilities to perceive. These systems should be able to use their knowledge developed while training on sub tasks (like understanding visual cues, generating text on the basis of this understanding) and use it while solving a larger, more complex task (like Identifying Affective State).

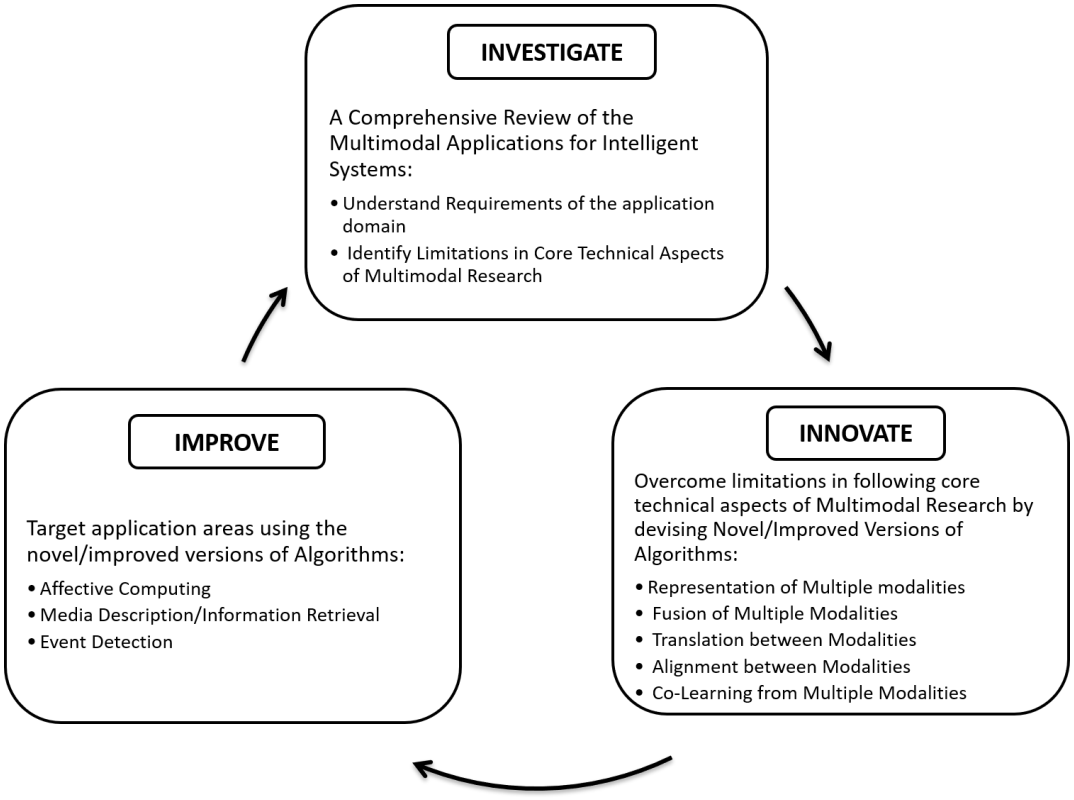
INVESTIGATE
A Comprehensive Review of the Multimodal Applications for Intelligent Systems:
• Understand Requirements of the application domain
• Identify Limitations in Core Technical Aspects of Multimodal Research
INNOVATE
Overcome limitations in following core technical aspects of Multimodal Research by devising Novel/Improved Versions of Algorithms:
• Representation of Multiple modalities
• Fusion of Multiple Modalities
• Translation between Modalities
• Alignment between Modalities
• Co-Learning from Multiple Modalities
IMPROVE
Target application areas using the novel/improved versions of Algorithms:
• Affective Computing
• Media Description/Information Retrieval
• Event Detection

Figure 1: Proposed Approach - Multimodal Modeling and Analysis of Intelligent Systems

Thus, the idea put forward here is to *Investigate* the limitations in existing approaches of core technical aspects, *Innovate* to overcome these limitations and apply/implement these innovations to bring about *Improvements* in the chosen set of Application Domains. This has been illustrated in Figure 1.

Proposed Research Objectives

- Investigate the core technical aspects of multimodal machine learning –
    - Fusion,
    - Representation,
    - Translation,
    - Alignment and
    - Co-Learning.
- Possible ways to improve the technical cores based on the limitations identified in the first objective.
- Implement improved versions of multimodal machine learning algorithms that employ perception capabilities to address unresolved problems in the application domains like Affective Computing, Media Description/Information Retrieval and Event Recognition.

It is proposed to limit the research in this thesis to an appropriate mixture and subset of the issues identified above.

References

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C. and Baskurt, A. (2011), "Sequential deep learning for human action recognition", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7065 LNCS, Springer, Berlin, Heidelberg, pp. 29–39.

Baltrušaitis, T., Ahuja, C. and Morency, L.-P. (2019), "Multimodal Machine Learning: A Survey and

Taxonomy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41 No. 2, pp. 1–1.

Bhandari, D. and Paul, S. (2018), "Perception Net: A Multimodal Deep Neural Network for Machine Perception", *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 1419–1426.

Borth, D., Breuel, T., Chen, T., Chang, S.-F. and Ji, R. (2013), "Large-scale visual sentiment ontology and detectors using adjective noun pairs", *Proceedings of the 21st ACM International Conference on Multimedia - MM '13*, ACM Press, New York, New York, USA, pp. 223–232.

Chao, L., Tao, J., Yang, M., Li, Y. and Wen, Z. (2015), "Long Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition", *ArXiv Preprint ArXiv:1212.5701*, pp. 65–72.

Chikersal, P., Poria, S., Cambria, E., Gelbukh, A. and Siong, C.E. (2015), "Modelling public sentiment in twitter: Using linguistic patterns to enhance supervised learning", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, available at:https://doi.org/10.1007/978-3-319-18117-2_4.

Dai, B., Fidler, S., Urtasun, R. and Lin, D. (2017), "Towards Diverse and Natural Image Descriptions via a Conditional GAN", *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2017–Octob, pp. 2989–2998.

Dhall, A., Goecke, R. and Gedeon, T. (2011), "Acted Facial Expressions In The Wild Database", *Technical Reports*, No. September, available at: http://cs.anu.edu.au/techreports/.

He, L., Jiang, D., Yang, L., Pei, E., Wu, P. and Sahli, H. (2015), "Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks", *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, No. October 2015, pp. 73–80.

Ji, S., Xu, W., Yang, M. and Yu, K. (2013), "3D Convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35 No. 1, pp. 221–231.

Karpathy, A., Joulin, A. and Li, F.F. (2014), "Deep Fragment Embeddings for Bidirectional Image

Sentence Mapping", *Advances in Neural Information Processing Systems*, pp. 1889–1897.

Kim, Y. (2014), "Convolutional Neural Networks for Sentence Classification", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1746–1751.

Kim, Y., Jernite, Y., Sontag, D. and Rush, A.M. (2015), "Character-Aware Neural Language Models", *ArXiv:1508.06615*, available at:https://doi.org/10.1002/mrm.1910060106.

Kiros, R., Salakhutdinov, R. and Zemel, R. (2014a), "Multimodal neural language models", *ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning -Volume 32*, Beijing, China, pp. 595–603.

Kiros, R., Salakhutdinov, R. and Zemel, R.S. (2014b), "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", *ArXiv: 1411.2539*, available at: http://arxiv.org/abs/1411.2539 (accessed 10 January 2019).

Kong, Y. and Fu, Y. (2018), "Human Action Recognition and Prediction: A Survey", *ArXiv:1806.11230*, pp. 23–48.

Lewkowicz, D.J. (2000), "The Development of Intersensory Temporal Perception : An Epigenetic Systems / Limitations View", *Psychological Bulletin*, Vol. 126 No. 2, pp. 281–308.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., et al. (2014), "Microsoft COCO: Common Objects in Context", *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, vol 8693 Springer, Cham, pp. 740–755.

Mahasseni, B. and Todorovic, S. (2016), "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 3054–3062.

Mihalcea, R., Morency, L. and Science, C. (2013), "Utterance-Level Multimodal Sentiment Analysis", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 973–982.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011), "Multimodal Deep Learning", *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696.

Niu, Y., Lu, Z., Wen, J.R., Xiang, T. and Chang, S.F. (2018), "Multi-Modal Multi-Scale Deep

Learning for Large-Scale Image Annotation", *IEEE Transactions on Image Processing*, IEEE, Vol. 28 No. 4, pp. 1720–1731.

Park, S., Shim, H.S., Chatterjee, M., Sagae, K. and Morency, L.-P. (2014), "Computational Analysis of Persuasiveness in Social Multimedia", *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, ACM Press, New York, New York, USA, pp. 50–57.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017), "A review of affective computing : From unimodal analysis to multimodal fusion", *Information Fusion*, Elsevier B.V., Vol. 37, pp. 98–125.

Poria, S., Cambria, E. and Gelbukh, A. (2015), "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, No. September, pp. 2539–2544.

Poria, S., Cambria, E. and Hazarika, D. (2017), "Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis", *IEEE International Conference on Data Mining (ICDM)*.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Hussain, A. and Gelbukh, A. (2018), "Multimodal Sentiment Analysis: Addressing Key Issues and Setting up Baselines", *ArXiv Preprint ArXiv:1803.07427*, available at: http://arxiv.org/abs/1803.07427.

Poria, S., Majumder, N., Morency, L.-P., Zadeh, A., Cambria, E. and Hazarika, D. (2017), "Context-Dependent Sentiment Analysis in User-Generated Videos", *ACL 2017- 55th Annual Meeting of the Association for Computational Linguistics*, pp. 873–883.

Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., Wood, I., et al. (2012), "Multimodal Emotion Recognition for AVEC 2016 Challenge", *ArXiv Preprint ArXiv:1212.5701*, pp. 75–82.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., et al. (2002), "Multimodal human discourse: gesture and speech", *ACM Transactions on Computer-Human Interaction*, Vol. 9 No. 3, pp. 171–193.

Ren, Z., Wang, X., Zhang, N., Lv, X. and Li, L.-J. (2017), "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward", *2017 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR), IEEE, pp. 1151–1159.

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J. and Goel, V. (2017), "Self-Critical Sequence Training for Image Captioning", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 1179–1195.

Richard Socher, Alex Perelygin, Jean Y.Wu, Jason Chuang, Christopher D. Manning, A.Y.N. and C.P. (2013), "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", *PLoS ONE*, available at:https://doi.org/10.1371/journal.pone.0073791.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., et al. (2015), "Av+Ec 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data", *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, pp. 3–8.

Ringeval, F., Sonderegger, A., Sauer, J. and Lalanne, D. (2013), "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions", *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, No. i, available at:https://doi.org/10.1109/FG.2013.6553805.

Rochan, M., Ye, L. and Wang, Y. (2018), "Video summarization using fully convolutional sequence networks", *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, Vol. 11216 LNCS, Springer, Cham, pp. 358–374.

Rohrbach, A., Rohrbach, M., Tandon, N. and Schiele, B. (2015), "A dataset for Movie Description", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 07–12–June, pp. 3202–3212.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., et al. (2017), "Movie Description", *International Journal of Computer Vision*, Vol. 123 No. 1, pp. 94–120.

Rouast, P. V., Adam, M. and Chiong, R. (2019), "Deep Learning for Human Affect Recognition: Insights and New Developments", *IEEE Transactions on Affective Computing*, pp. 1–1.

Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M. and Schiele, B. (2017), "Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training", *Proceedings of the*

IEEE International Conference on Computer Vision, Vol. 2017–Octob, IEEE, pp. 4155–4164.

Singh, B., Marks, T.K., Jones, M., Tuzel, O. and Shao, M. (2016), "A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 1961–1970.

Smith, L. and Gasser, M. (2005), "The Development of Embodied Cognition: Six Lessons from Babies", *Artificial Life*, Vol. 11 No. 1–2, pp. 13–29.

Torabi, A., Pal, C., Larochelle, H. and Courville, A. (2015), "Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research", available at: https://arxiv.org/pdf/1503.01070.pdf (accessed 16 January 2019).

Uzan, L. and Wolf, L. (2015), "I know that voice: Identifying the voice actor behind the voice", *Proceedings of 2015 International Conference on Biometrics, ICB 2015*, pp. 46–51.

Valstar, M., Pantic, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., et al. (2016), "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge", *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 3–10.

Varol, G., Laptev, I. and Schmid, C. (2018), "Long-Term Temporal Convolutions for Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40 No. 6, pp. 1510–1517.

Veeriah, V., Zhuang, N. and Qi, G.-J. (2015), "Differential Recurrent Neural Networks for Action Recognition", *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, pp. 4041–4049.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2014), "Show and Tell: A Neural Image Caption Generator", *ArXiv 1411.4555*, available at: http://arxiv.org/abs/1411.4555 (accessed 10 January 2019).

Wang, W., Ooi, C., Yang, X., Zhang, D. and Zhuang, Y. (2014), "Effective Multi-Modal Retrieval based on Stacked Auto-Encoding", *Proceedings of the VLDB Endowment*, Vol. 7 No. 8, pp. 649–660.

Weinzaepfel, P., Harchaoui, Z. and Schmid, C. (2015), "Learning to Track for Spatio-Temporal

Action Localization", *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, pp. 3164–3172.

Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L.P. (2013), "You tube movie reviews: Sentiment analysis in an audio-visual context", *IEEE Intelligent Systems*, Vol. 28 No. 3, pp. 46–53.

Wu, Z., Sivadas, S., Tan, Y.K., Bin, M. and Goh, R.S.M. (2016), "Multi-Modal Hybrid Deep Neural Network for Speech Enhancement", *ArXiv Preprint ArXiv:1606.04750*, available at: http://arxiv.org/abs/1606.04750.

You, Q., Luo, J., Jin, H. and Yang, J. (2015), "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks", *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, p. Pages 381-388.

Zadeh, A. (2015), "Micro-opinion Sentiment Intensity Analysis and Summarization in Online Videos", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pp. 587–591.

Zadeh, A., Zellers, R., Pincus, E. and Morency, L.-P. (2016), "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos", *IEEE Intelligent Systems*, Vol. 31 No. 6, pp. 82–88.

Zhang, S., Zhang, S., Huang, T. and Gao, W. (2016), "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition", *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*, pp. 281–284.